Neurostimulation reveals context-dependent arbitration between model-based and model-free reinforcement learning

Weissengruber, Sebastian ; Lee, Sang Wan ; O'Doherty, John P ; Ruff, Christian C

Abstract: While it is established that humans use model-based (MB) and model-free (MF) reinforcement learning in a complementary fashion, much less is known about how the brain determines which of these systems should control behavior at any given moment. Here we provide causal evidence for a neural mechanism that acts as a context-dependent arbitrator between both systems. We applied excitatory and inhibitory transcranial direct current stimulation over a region of the left ventrolateral prefrontal cortex previously found to encode the reliability of both learning systems. The opposing neural interventions resulted in a bidirectional shift of control between MB and MF learning. Stimulation also affected the sensitivity of the arbitration mechanism itself, as it changed how often subjects switched between the dominant system over time. Both of these effects depended on varying task contexts that either favored MB or MF control, indicating that this arbitration mechanism is not context-invariant but flexibly incorporates information about current environmental demands.

DOI: https://doi.org/10.1093/cercor/bhz019

Neurostimulation reveals context-dependent arbitration between model-based and model-free reinforcement learning

Sebastian Weissengruber [1] *, Sang Wan Lee [2] *, John P. O'Doherty [3], Christian C. Ruff [1]

* Co-first author

[1] Laboratory for Social and Neural Systems Research (SNS Lab), Department of Economics, University of Zurich, Zurich, Zurich, 8006, Switzerland

[2] Department of Bio and Brain Engineering, KAIST Institute for Artificial Intelligence & KAIST Institute for Health Science and Technology, KAIST, Daejeon, 34141, Republic of Korea

[3] Computation and Neural Systems Program & Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, California, 91125, USA

Running tilte:
Neural arbitration between model-based and model-free learning

Contact information:
Sebastian Weissengruber
University of Zurich, Department of Economics, Blümlisalpstrasse 10, 8006 Zurich, Switzerland
sebastian.weissengruber@econ.uzh.ch
+41446345037

Abstract

While it is established that humans use model-based and model-free reinforcement learning in a complementary fashion, much less is known about how the brain determines which of these systems should control behavior at any given moment. Here we provide causal evidence for a neural mechanism that acts as a context-dependent arbitrator between both systems. We applied excitatory and inhibitory transcranial direct current stimulation over a region of the left ventrolateral prefrontal cortex previously found to encode the reliability of both learning systems. The opposing neural interventions resulted in a bidirectional shift of control between model-based and model-free learning. Stimulation also affected the sensitivity of the arbitration mechanism itself, as it changed how often subjects switched between the dominant system over time. Both of these effects depended on varying task contexts that either favored model-based or model-free control, indicating that this arbitration mechanism is not context-invariant but flexibly incorporates information about current environmental demands.

Keywords: tDCS, reinforcement learning, ventrolateral PFC, goal-directed, habitual

Introduction

Optimal control of behavior requires humans and other animals to draw on their prior experiences with similar situations (Thorndike, 1933). How individuals use their experience to avoid punishment and maximize reward is well captured by reinforcement learning algorithms (Sutton and Barto, 1998). At least two fundamentally different strategies have been proposed in this respect (Daw et al., 2005): In model-free (MF) learning, past reward outcomes lead to an increased probability to repeat associated actions whereas in model-based (MB) learning, a model of the contingencies between states of the world, actions,

and outcomes is updated to guide choice. The more habitual MF learning therefore constitutes a simple “trial and error” strategy in which action values are updated by reward predictions errors that represent the discrepancy between expected and received reward. In contrast, MB learning is characterized by building and navigating through a model of the environment (that is, a “cognitive map”) that represents possible states, transition probabilities and outcomes. In this context, decisions are made by goal-directed planning and action values are updated by state prediction errors that represent the discrepancy between the environment and the internal model of it (for a more detailed overview see Dayan and Niv, 2008).

Understanding these two learning strategies and their neurobiological implementation is crucial not only for theoretical models of how humans make decisions but also for more applied purposes. This is because imbalances in the related psychological constructs of goal-directed versus habitual decision making have for a long time been associated with various psychopathologies, including drug abuse (Everitt and Robbins, 2005), obsessive-compulsive disorder (Gillan and Robbins, 2014) and Parkinson’s disease (de Wit et al., 2011; Redgrave et al., 2010). Computational algorithms of MB and MF learning offer a mathematical framework for studying the brain processes underlying goal-directed and habitual decision making (for a historical overview see, Dolan & Dayan, 2013). How tightly these constructs are linked is demonstrated by the finding that people who engage more in MB learning are also more sensitive to outcome devaluation (Gillan et al., 2015), a method commonly employed to measure goal-directed decision making (Adams and Dickinson, 1981; Balleine and Dickinson, 1998; Tricomi et al., 2009). More recent studies have also started to directly link the computational implementations of MB and MF learning to dysfunctional behavior in obsessive-compulsive disorder, binge eating disorder, and drug abuse (Voon et al., 2015), including alcohol dependence (Sebold et al., 2014). Diagnosis and treatment of such conditions may therefore benefit from a detailed understanding of

the neural mechanisms mediating MB versus MF control and particularly from demonstrations how these could be modified by external interventions.

Several studies (Deserno et al., 2015; Haruno and Kawato, 2006; Prévost et al., 2013; Wunderlich et al., 2012a) have therefore employed functional magnetic resonance imaging (fMRI) to investigate the neural implementation of MB and MF learning. These studies have revealed dissociable as well as common neural representations of both systems (for an overview, see O'Doherty et al., 2015). Moreover, several investigations have also addressed the question how these two neural subsystems may interact to drive choice. For instances, Gläscher and colleagues (2010) showed that behavior was best explained by a hybrid model that uses a combined action value derived from the weighted sum of MB and MF learning operations. Daw and colleagues (2011) detected areas in the striatum as well as the medial prefrontal cortex in which activity was correlated with both MB and MF prediction errors. Finally, Lee and colleagues (2014) found that activity in lateral prefrontal cortex correlated with the estimated reliability of both MB and MF learning systems. All these studies therefore suggest that dedicated neural mechanisms integrate the information provided by both the MB and MF system and possibly arbitrate between the two.

However, these neuroimaging studies alone leave it unclear which mechanism is causally involved in governing how the two learning systems interact to drive behavior. This is because neuroimaging techniques only reveal correlations between model predictions and neural activity, which by themselves are not informative about whether the correlated neural activity causally drives the choices or only reflects a functionally irrelevant byproduct that arises as a consequence of behavior. This limitation is aggravated by the fact that the same behavioral data can be fit by very different computational models, the predictions of which may correlate with different spatial patterns of neural activity (Gläscher and O'Doherty, 2010; Mars et al., 2012; O'Doherty et al., 2007). The same

neuroimaging data may therefore suggest rather different underlying neural mechanisms, depending on which model is fit to the data. Overcoming these limitations requires methods that can test how neural activity in circumscribed brain areas causally contributes to observable behavioral changes.

In one study employing such methods, Wunderlich and colleagues (2012b) were able to increase MB behavior in a behavioral task with a dopamine agonist. However, while this finding may be valuable for clinical purposes, the neural mechanisms underlying the pharmacological effects remain unclear. In another set of studies, Smittenaar and colleagues attempted to modulate MB and MF learning via transcranial magnetic stimulation (TMS) (Smittenaar et al., 2013) and transcranial direct current stimulation (tDCS) (Smittenaar et al., 2014). While the latter study did not lead to any effects on behavior, reduced MB behavior was observed after repetitive TMS (Smittenaar et al., 2013) over a dorsolateral prefrontal cortex (dlPFC) region initially defined based on its role in working memory (Feredoes et al., 2011). The strength of this disruptive effect on MB learning indeed depended on the individual visuspatial working memory capacity of subjects (Smittenaar et al., 2013). Thus, this result concurs with other findings (Otto et al., 2013) that well-functioning working memory is more important for MB then MF learning, presumably reflecting the demand for storing a cognitive map of reward contingencies. However, these findings do not reveal whether the stimulated dlPFC area indeed implements neural mechanisms that integrate or arbitrate between the MB and MF systems during reinforcement learning. The neural mechanisms that allow humans to flexibly shift between MB and MF control depending on environmental demands are therefore still unknown.

Here we addressed this issue by investigating an arbitration mechanism recently proposed by Lee and colleagues (2014). Their computational model assumes that agents weigh MB and MF learning strategies based on each system’s current reliability, as defined by the

relationship of reward- and state prediction errors over time. Signals indicating the reliability of the dominant system at any moment of time were found to correlate with neural activity in an area of the ventrolateral prefrontal cortex (vlPFC), suggesting that this area may contain an arbitration mechanism that flexibly selects the system most suited for the current control of behavior. Connectivity analyses further suggest that this prefrontal area might arbitrate between both systems by inhibiting striatal areas involved in MF processing if the MB system is deemed more reliable (Lee et al., 2014). In the present study, we directly tested the causal relevance and functional properties of this proposed mechanism, by examining how enhancing or reducing neural activity via neurostimulation affects the dominance of, and arbitration between, the MB and MF systems during reinforcement learning.

We chose tDCS (Nitsche et al., 2008; Utz et al., 2010) as the method of neurostimulation due to its well-established polarity-specific effects on neural excitability: If the stimulation is applied within defined parameter ranges and contexts (Batsikadze et al., 2013; Woods et al., 2016), anodal stimulation causes depolarization and higher excitability of neurons, whereas cathodal stimulation causes hyperpolarization and decreased excitability (Nitsche and Paulus, 2001, 2000; Nitsche et al., 2003). Both these types of tDCS are safe to apply (Poreisz et al., 2007) and have already been used successfully to modulate decision-making in other contexts (Fecteau et al., 2007; Hecht et al., 2010; Knoch et al., 2007; Ruff et al., 2013). By applying both polarities of tDCS over the left vlPFC area identified by Lee and colleagues (2014), we could therefore investigate the functional properties of the putative arbitration mechanism.

Any mechanism arbitrating between MB and MF learning not only has to control how much an agent relies on one or the other system, but also when and how much one switches between them in line with the current demands imposed by the environment. Hence, to fully clarify the neural and behavioral relevance of the targeted mechanism in

reinforcement learning, we applied the stimulation during a learning paradigm with varying environmental demands. This was achieved by using two types of task periods that either favored MB or MF control, as they differed in the necessity to use a model of transition contingencies for successful task performance. In both of these task conditions, we assessed if arbitration between learning systems was executed by either implementing a static bias for one or the other system, by changing the temporal dynamics of switching between systems, or by a combination of both these strategies. If the targeted mechanism is indeed involved in one or both of these arbitration strategies, then up- or down-regulating its function should lead to corresponding effects on behavior.

More specifically, in task periods that force subjects to make MB decisions to obtain rewards, successful arbitration should shift dominance towards the MB system. Since the arbitrator has been proposed to achieve this by inhibiting the MF system (Lee et al., 2014), enhanced neural excitability of the arbitration mechanism due to anodal tDCS should therefore lead to stronger MF inhibition and favor MB processing, while decreased neural excitability due to cathodal tDCS should disinhibit the MF system and thus render it more dominant.

However, in task periods for which subjects do not necessarily require a model of the environment to perform well, implementing a static bias for MB learning would not constitute an optimal strategy. In order to perform well in these contexts, subjects should rather find an appropriate equilibrium between MB and MF control. Thus, the arbitrator should be more likely to balance between systems, rather then shifting the dominance in favor of one. To test this conjecture, we investigated whether tDCS would lead to changes in the frequency of the switches between both learning systems.

Materials and Methods

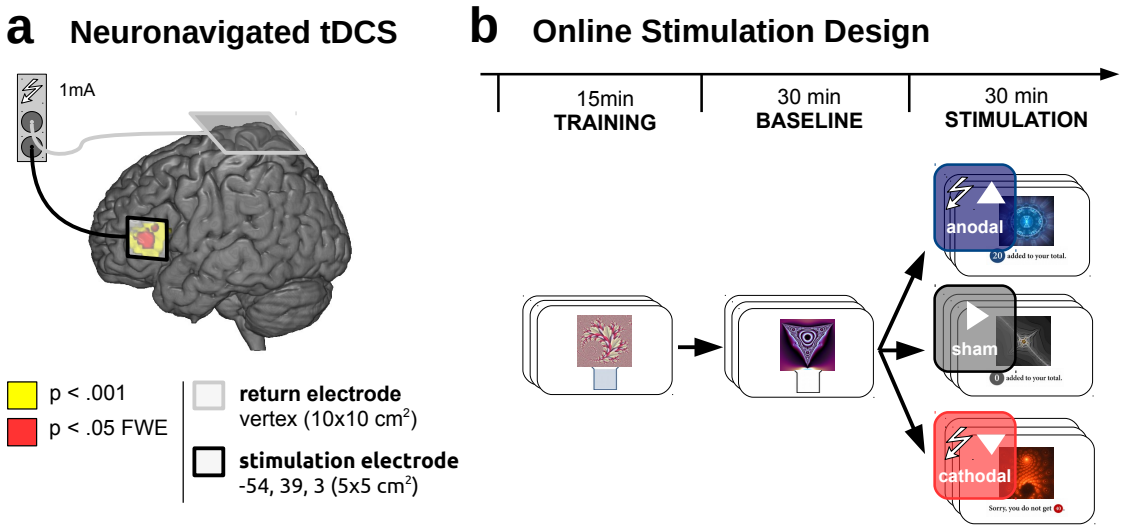
Participants and Study Design

To examine how the application of MB and MF learning within this task was governed by the activity reported by Lee and colleagues (2014), we applied tDCS over the vlPFC area identified in that study (Figure 1a). Based on our prior experience with neurostimulation and the procedures of comparable studies (Smittenaar et al., 2014), we recruited 60 right-handed, healthy subjects (out of a Swiss student volunteer pool) and randomly assigned them to receive neuronavigated tDCS using either an excitatory anodal, suppressive cathodal, or sham stimulation protocol. Subject attrition for the stimulation sessions, paired with the necessity for individual structural fMRI images used by our neuronavigation approach (see Neurostimulation section), resulted in a final sample 55 subjects (18 anodal, 20 cathodal, 17 sham; 21 female; mean age = 23.07, SD = 3.29). All subjects were screened for physiological suitability to tDCS via an initial telephone interview and on-site standardized questionnaires, were given the chance to clarify any open questions, reported no history of psychiatric or neurological disease, and gave informed written consent to the experiment. All experimental procedures were approved by the ethics committee of the canton of Zurich (Kantonale Ethikkommission Zürich).

Each subject performed a training session and two consecutive experimental sessions of the task. The first session was conducted without stimulation and constituted a baseline measure, whereas the second session was conducted while we applied the respective stimulation protocol (Figure 1b). This allowed us to examine how the stimulation changed each participant's individual learning style in a mixed-design difference-in-difference analysis (Baayen et al., 2008). While subjects were waiting for the tDCS electrodes to be attached, they were instructed to undergo a short working memory test (n-back task; progressing from 1-back to 4-back) adapted from Gevins and Cutillo (1993). An analysis of

the 51 subjects who successfully completed this task (16 anodal, 19 cathodal, 16 sham; 4 subjects were excluded due to a misunderstanding of the task instructions) confirmed that the three stimulation groups did not significantly differ in short term memory capacity (anodal: M = 9.75, SD = 5.04; sham: M = 9.81, SD = 3.80; cathodal: M = 11.79, SD = 2.59; $F(2,48) = 1.61$, $p = .21$). The participants then read the instructions of the two-choice decision task (available as supplementary material) and had the opportunity to ask questions. Once the electrodes were applied, participants underwent a short training session followed by the two consecutive sessions of the task (baseline and active stimulation or sham depending on the subject's experimental group) lasting approximately 25 to 30 minutes each (Figure 1b). Participants were told that they will be paid 20 Swiss francs plus the outcomes of two random trials for each session (see Task and Stimuli section). These payoffs were calculated in a way that a participant with average performance level would receive roughly 100 CHF (approximately 105 USD at the time of testing). In order to make the payment of subjects transparent, we used an adapted version of the Prince incentive system (available at the Social Science Research Network (SSRN), 2504745). Before the start of the task, each participant received a sealed envelope containing 4 random trial numbers. After subjects finished the task, they were presented with a table of all their choices in both sessions and opened the letter to see which of their choices would be rewarded. This approach assured that subjects could be sure that trials were really selected randomly, while creating the same incentive for each trial without any dependency on prior choice. All 55 participants that took part in the study finished the main experiment and were included in the final analysis.

Figure 1. Experimental Design and Procedures



(a) *Area targeted with tDCS based on fMRI results of Lee et al. (2014). We defined the stimulation site as the lPFC area hypothesized to arbitrate between learning systems: Colored areas represent voxels that covary with the reliability signal (based on state and reward prediction errors) of the currently dominant reinforcement learning system. Rectangles represent the position of tDCS sponge electrodes over the stimulation and return site.*

(b) *Depiction of the experimental design. Each subject went through the same sequence of task sessions but was randomly assigned to one of three stimulation conditions: anodal, sham or cathodal. The first session was intended as a baseline. The three different types of online stimulation were applied during the whole second session of the task.*

Neurostimulation

Stimulation coordinates in MNI space were defined as x = -54, y = 39, z = 3, corresponding to the peak voxel that covaried with the reliability signal of the currently dominant reinforcement learning system in the study of Lee and colleagues (2014) (Figure 1a). To use accurate neuronavigation, we segmented the grey matter of T1-weighted structural scans for all subjects and normalized the MNI coordinates to their individual

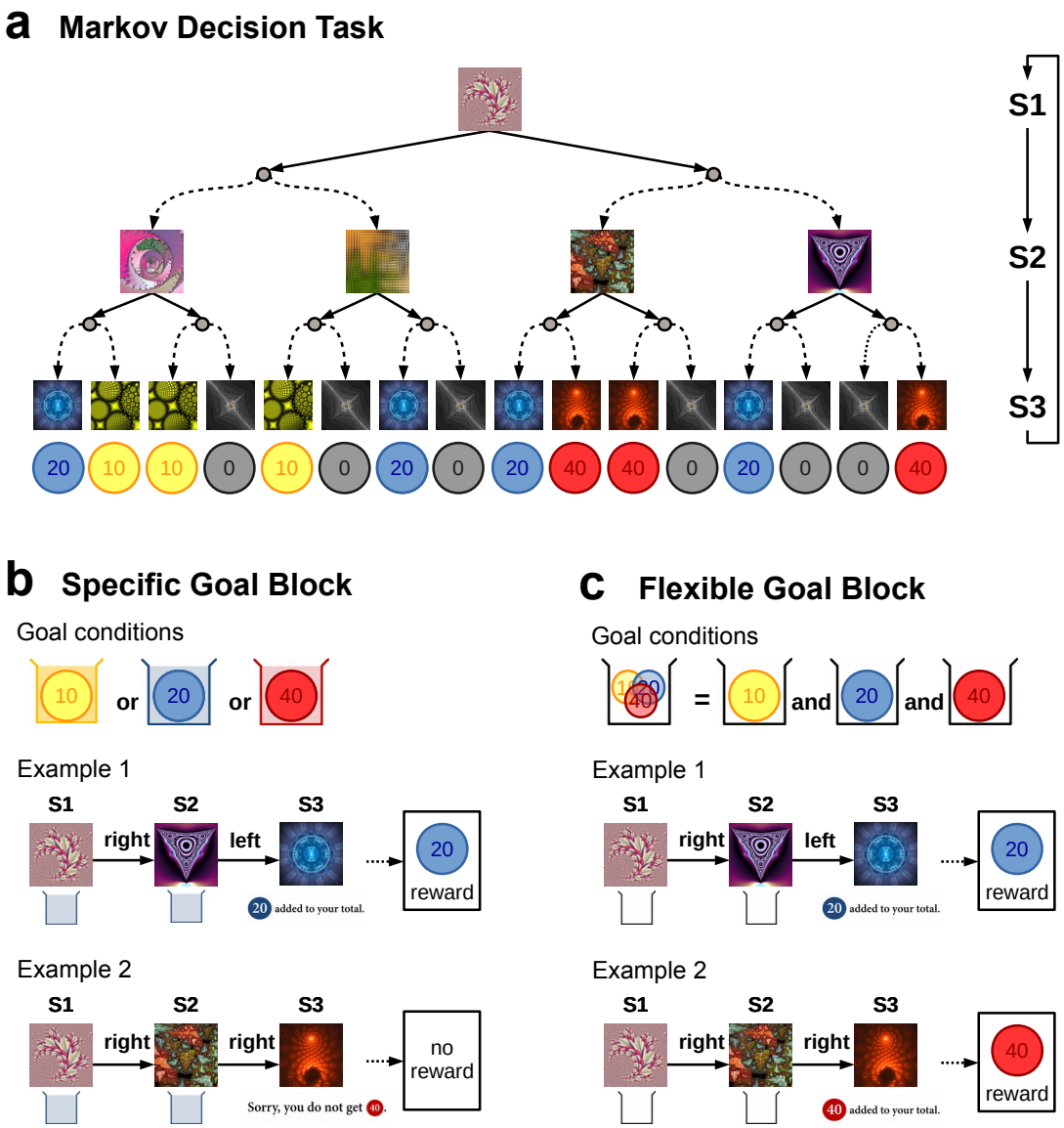
brain space. We then used Brainsight, a guided stereotactic neuronavigtion system (Rogue Resolutions Ltd., Cardiff, UK), to create curvilinear 3D reconstructions of subjects' brains and mark the electrode position as cranial projection of individual coordinates tangentially to the cortex. Subsequent electrical stimulation was applied with a 16 channel DC-STIMULATOR MC (neuroConn GmbH., Ilmenau, Germany) and two saline water soaked sponge electrodes fixed with rubber bands. We used a 5x5 cm^2 electrode at the target location and a large 10x10 cm^2 electrode over the vertex. This strategy minimized any neural effects under the vertex electrode (Nitsche et al., 2008). Active stimulation was performed during the full second session of the task, including a 5-minute resting period before the task to allow for stabilization of the effects. We applied constant direct current of 1mA for a maximum of 30 minutes, with ramp-up and ramp-down times of 25 seconds each. Sham stimulation mimicked the physical sensation of active tDCS, by applying the identical current strength with the same ramp-up time, followed by 25 seconds of stimulation plus the same ramp-down time. Impedance was kept below 20 kΩ for all subjects resulting in voltage below 20V. The computer controlling the stimulation protocol and the randomization of subjects was preprogrammed and in a different room, so that neither the subjects nor the experimenters attaching electrodes or marking their position had information about the stimulation condition. Except for the direction of current, the tDCS montage itself was identical for the sham, anodal, and cathodal groups.

Task and Stimuli

In order to investigate the causal role of the neural signals reported by Lee and colleagues, we employed the same task as in their original neuroimaging study (Lee et al., 2014). Participants were given the chance to collect rewards by navigating through the states (represented by fractal images) of a two-choice, three-stage probabilistic decision tree. This required participants to make two sequential decisions between different fractals

in order to reach the final state that was associated with reward outcomes, as represented by colored coins (Figure 2a). Thus, similar to other two-choice Markov decision tasks, participants could solve the task by deciding mainly based on the reward history associated with different actions at each stage (MF learning) or by constructing and navigating a complex model of the associations between different fractal images and their associated reward (MB learning). To introduce environmental demands that either favored MB or MF control, we used two task conditions with qualitatively different goals. In the “specific” goal condition, only coins of a specific color could be collected. This encouraged MB processing because subjects had to build a “cognitive map” to successfully navigate to a specific final state (Figure 2b). In flexible trials, all coins could be collected. This allowed subjects to use both learning strategies, but encouraged the simpler MF processing, as reinforcement of positive outcome decisions was sufficient for good performance (Figure 2c).

Figure 2. Task Design



(a) *Schematic decision tree of a subject with the three stages S1 – S3. In any given trial, subjects always started in the same state (at stage S1) and had to make a decision ("left" or "right") which probabilistically led them to one of two possible states at the next stage (S2). After a second decision in S2, the subject reached the final state (at S3) and collected the assigned coin associated with monetary reward. Solid, straight arrows represent possible subject decisions (left, right) and dotted, curved arrows represent probabilistic transitions.*

(b) *Examples of possible trial sequences in the "specific" condition that favored MB*

processing. On the screen underneath the fractal representing the current state, a collection box always displayed the color of the coins subjects were able to collect on this round. If subjects successfully navigated to the correct state, they received the coin of the rewarded color (as in Example 1). If the coin resulting from the final choice did not match the correct color, subjects did not receive anything (as in Example 2).

***(c)** Examples of trials in the "flexible" condition that favored MF processing. The white collection box indicated that coins of all colors could be collected. Any colored coin that participants reached after two choices was added to their winnings (as in Examples 1 and 2).*

The task was composed of 9 different possible states, represented by fractal images, which we randomized for each subject to create an individual decision tree (as exemplified in Figure 2a). Each tree featured a starting point (S1) and two consecutive stages (S2, S3) that could adopt different states. Participants were not aware of the structure of the task and were only presented with the respective fractal of a stage. They had to navigate through the states by pressing the left or right arrow on a keyboard. After starting with the same state (S1) on each trial they had to make a decision ("left" or "right") that led them to a specific state of the next stage (S2) via a probabilistic state-transition. In one trial subjects always made two consecutive decisions in order to reach a final stage (S3) for which each state was associated with a colored coin. After subjects received (or did not receive) a coin, the next trial started at the old starting point (S1). Once such a tree was defined, the possible states stayed constant for the rest of the experiment. This enabled subjects to learn the tree's possible transitions and exploit them to collect the desired coins. Subject had four seconds to decide after they were presented with their current state and an indicator of which coins they were able to collect (Figure 2b, c). Once they made a decision, the next state was presented after 150 milliseconds. When subjects

reached the third and last stage, the coin indicator disappeared and was replaced with the actual coin associated to the state for 2 seconds. A sentence “[COIN] added to your total.” or “Sorry, you do not get [COIN].” was shown accordingly (Figure 2b).

Each of the two task sessions (baseline and stimulation) consisted of 56 blocks. Each of these blocks could differ in goal condition (flexible or specific, indicated to the subject, see Figure 2b, c) and also in transition probabilities (low and high state-transition uncertainty, which was not indicated to the subjects). Trials in “low uncertainty” blocks had a 90% probability to reach the more likely state, while the “high uncertainty” blocks featured a 50% chance to get one or the other possible next state. Blocks with low state-transition uncertainty lasted for 3 to 5 trials in a row, while high uncertainty blocks were designed to last for 5 to 7 trials. In each session, the 56 blocks (14 of each type, flexible or specific paired with either low or high uncertainty) were ordered in a randomized sequence, which added up to 280 trials per session on average. Before the actual task sessions started, subject underwent a training session consisting of 80 flexible trials followed by 20 specific trials, to allow subjects to familiarize themselves with both conditions. The training trials were identical to the trials used in the two main sessions, but did now allow subject to earn any rewards. These trials were thus designed to allow subjects to start learning the task contingencies for the following sessions. This training, and the alternation of block length and state-transition probabilities in the main experiment, were used to facilitate meaningful learning while securing that the task was not too easy for the number of trials presented. The values listed above were confirmed to achieve that in a pilot experiment with different subjects recruited from the same population. In order to incentivize participants to use reinforcement learning to improve their performance rather than acting randomly, all collected coins had a chance to be converted to real money after the experiment (if the trial number of a collected coin was contained in the envelope, its value was paid out in CHF as described in the Participants and Study Design section). This nature of incentives

ensured that participants were motivated to use the best-working strategy on every trial, as all collected coins had a probability to be converted to real money after the experiment. The task was coded with Psychtoolbox (Brainard, 1997) in Matlab (The MathWorks, Inc., Natick, Massachusetts, USA).

Computational Modeling

To quantify the effect of the stimulation on our subjects' learning behavior while avoiding possible interactions between the many internal parameters of an arbitration model, we independently fitted MB and MF reinforcement learning models to the individual choice data of our subjects, for both the baseline and stimulation session. This allowed us to generate a parameter that quantified the preference for MB versus MF learning, by comparing the likelihoods of the two learning systems for each trial. We also created a second parameter quantifying how much subjects switched between both learning systems by counting how often trials with higher likelihood of one system were followed by trials with higher likelihood of the other system.

To do so we implemented a model-free SARSA learner (Sutton and Barto, 1998) and a model-based learner (Gläscher et al., 2010; Lee et al., 2014). The MF and MB learners used reward prediction errors and state prediction errors to compute state-action values, respectively. In the former case, the learning process was driven by observations of rewards, whereas in the latter case, the learning process modified a model of the environment representing the state-action-state transition probabilities. Each model was then independently fitted to each session’s choice data for each individual, comprising 220 types of learners in total (MB and MF learner for baseline and stimulation session for 55 subjects). Next, we computed the likelihoods of the two learners for each trial. Comparison of these likelihoods allowed us to detect shifts in preference for MB over MF control on a trial-by-trial basis and to examine how the preference for MB over MF control is influenced

by the three stimulation types (anodal, sham or cathodal) in the two different goal conditions (specific and flexible).

In the model-free learner (Sutton and Barto, 1998) (MF), the amount of updates of the state-action value $Q_{MF}(s,a)$ for the action a in the state s is defined as the reward prediction error (RPE) δ_{RPE} :

$$\delta_{RPE} = r(s') + \gamma Q_{MF}(s',a') - Q_{MF}(s,a),$$
$$\Delta Q_{MF}(s,a) = \alpha \delta_{RPE},$$

where s, s' refers to the current and the next state, respectively, a, a' refers to the action in the current state and in the next state, respectively, $r(s')$ denotes the obtained reward in state s', γ is a temporal discount factor (Gläscher et al., 2010) fixed at 1, α denotes the learning rate, the model's free parameter.

In the model-based learner (MB), the update of the state-action value is performed through a combination of FORWARD learning and BACKWARD planning (Lee et al., 2014). The FORWARD learning component uses experience with state transitions to update the matrix $T(s,a,s')$ of state-transition probabilities, which represents the probability of the agent's state being s' if it made a choice a in a state s . Whenever the agent's state transition occurs, the state prediction error (SPE) is computed and the corresponding state-action value is updated:

$$\delta_{SPE} = 1 - T(s,a,s'),$$
$$\Delta T(s,a,s') = \eta \delta_{SPE},$$
$$Q_{MB}(s,a) = \sum_{s'} T(s,a,s') \{ r(s') + max_{a'} Q_{MB}(s',a') \},$$

where η denotes the learning rate, the model's free parameter. The first term of the SPE is set to 1 in order to incorporate that the state space is assumed to be deterministic.

The second component of the model-based learning is the BACKWARD planning (Lee et al., 2014). The agent goes through this process whenever it is presented with an explicit goal (e.g., change in a specific goal condition or transition from the flexible to the specific goal condition). To update the value of each state, the FORWARD update process is

repeated backwards for all possible states and actions:

$$r(s) = \begin{cases} R & for\ a\ goal\ state, \\ 0 & otherwise. \end{cases}$$

$$\begin{aligned} &for\ \ i = 3,2, \\ &\quad for\ \ s \in S_{i-1} \\ &\qquad Q_{MB}(s,a) = \sum_{s'} T(s,a,s')\{r(s_i) + max_{a'} Q_{MB}(s',a')\},\ \ for\ all\ a. \\ &\quad end \\ &end \end{aligned}$$

where R is the reward value of the goal state, S_i refers to the set of states in i-th stage.

Both the MB and the MF learner select actions stochastically according to the following softmax function (Gläscher et al., 2010; Luce, 1959):

$$P(s,a) = \frac{\exp(\tau Q(s,a))}{\sum_b \exp(\tau Q(s,b))} \quad ,$$

where τ is the inverse temperature parameter controlling the extent to which the agent made a choice with the higher valued action.

We used the Nelder-Mead simplex algorithm (Lagarias et al., 1998) to estimate the free parameters of the MB and the MF learners (the learning rate and the inverse temperature of the softmax function) by minimizing negative log-likelihood $-\sum \log(P(s,a))$ of the obtained choices given the observed choices and rewards, summed over all trials for each subject. To minimize the risk of finding a local but not global optimal solution, we ran optimization 100 times with randomly generated seed parameters.

Statistical Analysis

Effects for which we had a clear a-priori bidirectional hypothesis (namely tDCS effects on model preference, model switching, and choice switching) were tested on the coefficients of a linear regression model. We accounted for the proposed opposite directionality of anodal and cathodal stimulation by coding the tDCS groups as a contrast with weights 1, 0, -1 for anodal, sham and cathodal stimulation, respectively. These bidirectional

hypothesis test were followed by post-hoc analysis (two-sample t-tests). To test for interactions between stimulation condition, task conditions, and model parameters, we also included the latter two as categorical predictor variables in the model and added the corresponding interaction terms. To account for natural changes in behavior independent of stimulation, we conducted these analyses on the differences between baseline and stimulation session (calculated separately for the three stimulation conditions in each of the two task conditions). For other measurements of interest (for example performance parameters), we used a linear mixed-effects model, which allowed us to compare both types of stimulation as separate condition factors against sham, while accounting for the time factor between baseline and active sessions and treating subjects as random effects (different intercepts). For results that did not test stimulation effects (for example specific versus flexible trials within the baseline condition, correlations between model preference and reward, changes over time within the sham condition, or working memory capacity between groups) we used paired sample t-tests, Pearson's correlations and ANOVAs, respectively. All p-values were calculated using two-tailed testing except when noted explicitly as one-tailed. The statistical analysis was performed with the Matlab statistics toolbox (The MathWorks, Inc., Natick, Massachusetts, USA).

Results

As hypothesized, the different task conditions had a strong effect on learning strategies in the baseline session before neurostimulation was applied. In the specific condition, the MB learner was dominant in 61% of trials, whereas in the flexible condition, it provided a better fit for only 33% of trials ($t(54) = 15.87$, $p < .001$) (Figure 3a). Additionally, a stronger preference for MB learning was positively correlated with performance (mean points per trial) in specific blocks ($R(53) = .36$, $p = .007$) but correlated negatively with performance in flexible blocks ($R(53) = -.62$, $p < .001$) (Figure 3b). This confirms that specific and flexible

blocks indeed favored MB and MF processing, respectively. Switching between learning systems was also increased in specific (M = 53%, SD = 10.17) compared to flexible trials (M = 47%; SD = 8.93), ($t(54) = 2.46$, $p = .017$) but did not show any significant association to performance in the baseline session (specific, $R(53) = .15$ $p = .27$; flexible, $R(53) = -.08$, $p = .58$). Finally, we also tested for any possible changes of these parameters over time (irrespective of stimulation conditions). We did so by comparing the initial baseline and the subsequent experimental session within the sham group. This revealed that in the specific task condition, participants not only favored MB strategies (Figure 3a) but also further increased MB processing over time from 61% (SD = 9.55) to 65% (SD = 9.73) ($t(16) = 2.53$, $p = .023$). In the flexible condition, however, participants decreased their preference for MB control from 35% (SD = 8.23) to 32% (SD = 6.05) ($t(16) = -2.40$, $p = .029$). No such effects were found for the switching parameter between learning systems (specific, $t(16) = .35$, $p = .73$; flexible, $t(16) = 1.09$, $p = .29$). Subjects therefore not only favored a different strategy for each task condition, but further amplified their reliance on one or the other learning system over the course of the task. To account for these naturally occurring changes over time, we based the analysis of stimulation effects on the changes from baseline to stimulation session.

Behavioral Effects of Stimulation in the Specific Condition

To test our predictions about the functional role of the stimulated neural mechanism, we analyzed and compared how individual learning strategies (as defined by the model preference and model switching parameters) changed relative to baseline during the three types of tDCS, using a linear regression model. In the specific task condition that favored MB learning, subjects receiving anodal tDCS (thought to increase neural excitability) indeed exhibited enhanced MB learning whereas participants receiving cathodal tDCS (thought to decrease excitability) demonstrated more pronounced MF learning ($t(53) =$

2.75, p = .008; Figure 3c, left). Consistent with our hypothesis, this suggests that the stimulated mechanism may increase MB learning by inhibiting the MF system in circumstances where its predictions are less reliable. Additionally, we investigated how the arbitration mechanism between MB and MF learning was affected by analyzing how much subjects changed their frequency of switching between dominant learning systems during stimulation. However, we could not detect any effects indicating a change in model switching behavior between the different stimulation groups (t(53) = .12, p = .90; Figure 3c, right). Our study was designed to test the hypothesized bidirectional effects brought about by anodal and cathodal stimulation, rather than characterizing individual effects of either stimulation type. However, a post-hoc analysis of the differences between either active simulation condition and the sham stimulation further reveals that the effect on learning strategy preference differed significantly between the cathodal and sham control condition (cathodal vs. sham, t(35) = -1.81, p = .039; one-tailed) but not between the anodal and the sham condition (anodal vs. sham, t(33) = .92, p = .18; one-tailed).

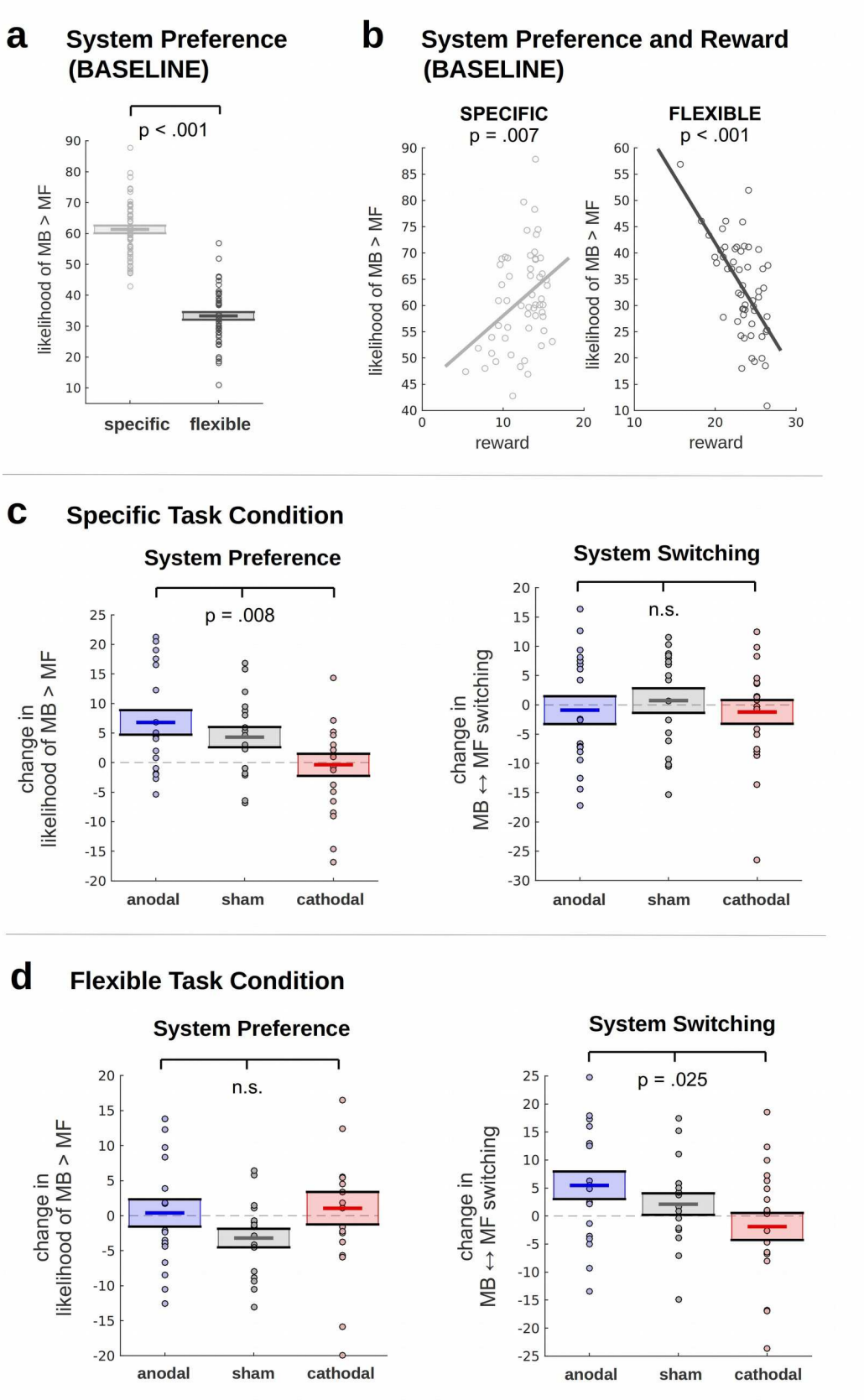
Behavioral Effects of Stimulation in the Flexible Condition

In the flexible task condition that allowed for both MB and MF decisions, but drove participants to behave significantly more MF than in specific trials (Figure 3a), we could not find any difference in tDCS effects on the preference for one over the other learning system (t(53 = -.29, p = .77; Figure 3d, left). However, subjects receiving anodal stimulation showed increased switching between dominant learning systems whereas the cathodal tDCS group showed reduced switching (t(53) = 2.31, p = .025; Figure 3d, right). This indicates that the affected neural mechanism is not merely engaging in a unilateral inhibition of MF control, but is able to selectively switch between learning models when both MB and MF systems are relevant.

We confirmed that these effects are indeed specifically expressed for just one or the other

parameter (preference or switching) in just the specific or flexible condition, respectively, by calculating the interaction between stimulation groups, model parameters and task conditions. This was significant ($t(212) = 2.59$, $p = .010$). Thus, the stimulated model arbitration mechanism exerts different behavioral effects in the flexible and specific task conditions, which differ in their demand to store a cognitive model of the decision tree. This supports the idea that the stimulated control mechanism adapts itself to the demands of the current environment (see Discussion section). Post-hoc tests showed that the tDCS effects on learning-strategy switching behavior for cathodal and anodal tDCS alone, compared to sham stimulation, just failed statistical significance (anodal vs. sham, $t(33) = 1.06$, $p = .15$; cathodal vs. sham, $t(35) = -1.26$, $p = .11$; one-tailed). Thus, we cannot perform statistical inference on the effectiveness of either stimulation protocol considered in isolation.

Figure 3. Behavioral Effects of tDCS



(a) *Preference for MB control (% of trials where the MB learner has a higher likelihood than the MF learner) between task conditions during the baseline session. The analysis of*

the derived preference parameter confirmed that subjects engaged more in MB learning in the specific task condition but predominantly used a MF strategy in the flexible condition.

(b) *Preference for MB control and performance (average points earned per trial) correlated positively in specific blocks, but negatively in flexible blocks. This result demonstrates that the two task conditions indeed favored different learning strategies.*

(c,d) *Differences in the change from baseline of derived model parameters between stimulation groups. tDCS affects either learning system preference or learning system switching (% of trials where the dominant system changed between MB or MF), depending on environmental demands. In specific trials favoring MB control (c) only MB preference is strengthened/weakened by anodal/cathodal tDCS. In flexible trial blocks favoring MF control (d) only model switching is increased/decreased by anodal/cathodal tDCS. Note that these effects are indeed specific to a combination of parameter, task condition, and tDCS group, as indicated by a significant three-way interaction of these factors ($t(212) = 2.59$, $p = .010$).*

In all plots, *circles represent individual data points. If applicable, the central (colored) mark is the mean, the bottom and top of the shaded area represents the standard error of the mean (SEM).*

Stimulation Effects on Task Performance

Since we showed that tDCS led to a bias in preference for one or the other learning system, and that this preference was related to the earnings of subjects in the pre-stimulation baseline session (see Figure 3b), we examined how tDCS affected objective task performance. We did so by testing how the different stimulation conditions changed choice optimality (defined as percentage of decisions that were optimal for reaching the highest expected value on any trial) and task performance (collected rewards), relative to each participant's baseline. Cathodal stimulation indeed resulted in reduced optimality

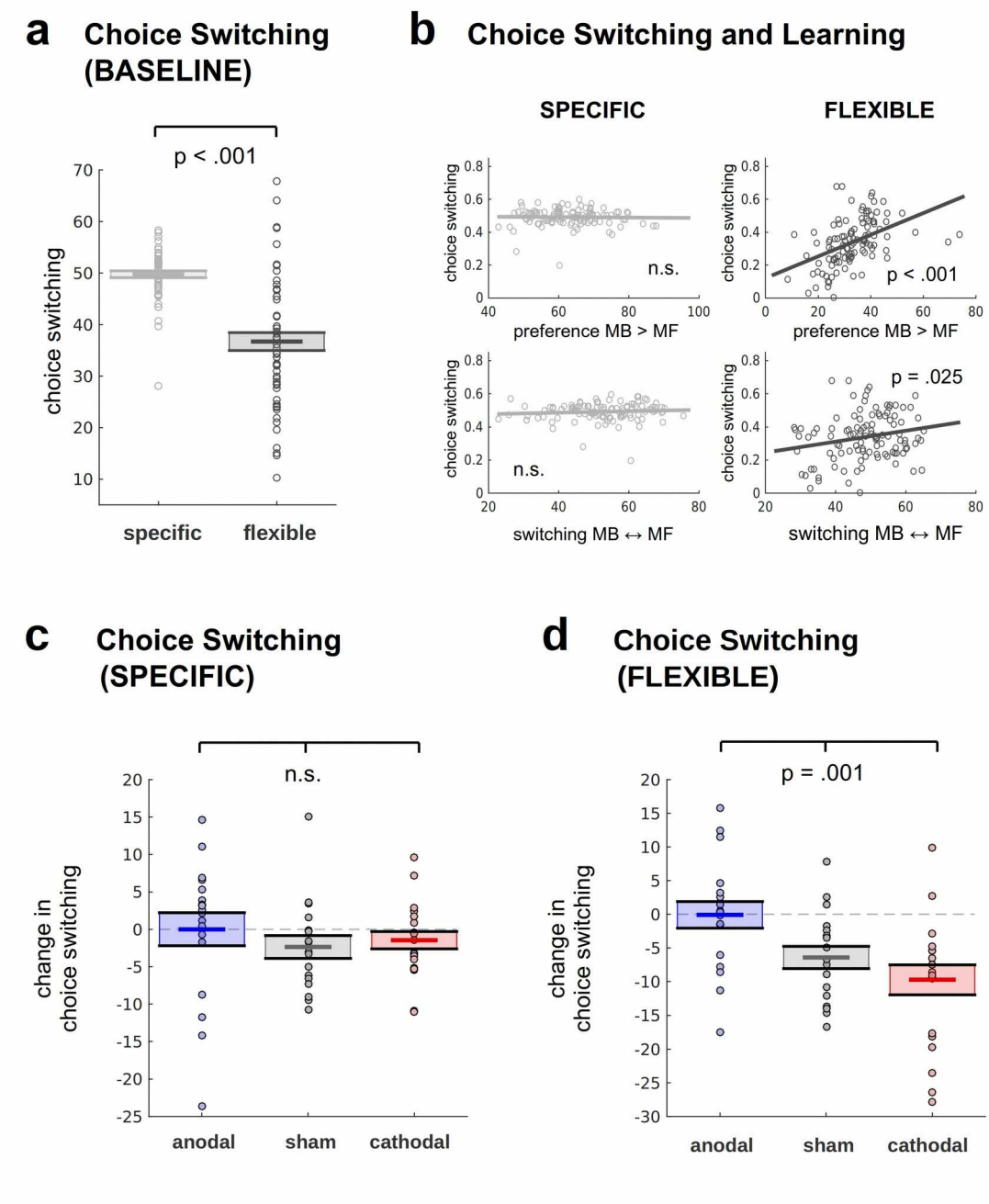
(F(1,55) = 4.29, p = .043) and lower objective performance (points per trial) (F(1,55) = 4.54, p = .038) in the specific condition, consistent with the fact that cathodal tDCS decreased MB learning on specific trials where MB control was crucial for (and correlated with) higher performance. By contrast, anodal stimulation neither led to enhanced decision optimality (F(1,55) < .001, p = .99) nor to a concomitant increase in task performance in the specific task condition (F(1,55) = 0.14, p = .71). However, this is likely to reflect a ceiling effect: Subjects generally showed high optimality in the task (choosing the action with the highest expected value in 82% (SD = 9.96) of cases) and also had to explore alternatives with lower expected value to learn about the state structure and contingencies of the task, thereby limiting the possibility of tDCS to further enhance choice optimality.

Non-algorithmic Behavioral Effects of Stimulation

In order to rule out that the behavioral effects caused by tDCS are only expressed for the specific learning model we fitted to the data, we also conducted a simpler, non-algorithmic analysis. To do so, we inspected choice switching, a simple count of how often subjects systematically repeat or switch between actions from one trial to the next. We used this measure as an alternative to capture shifts in learning strategies independently of the assumptions of computational algorithms. To calculate an index for this simple type of choice switching, we counted how often subjects switched away from their previous choice in the second decision stage (S2). Since the choice of S2 is contingent on the choice in S1, this index represents a combined metric of choice consistency in both stages. Our task design implied that subjects should react in their choices more to the changing goals of specific trials. Behavior in the baseline session confirmed this feature of our design: choice switching was indeed considerably higher in specific than flexible trials (t(54) = 7.93, p < .001) (Figure 4a). Crucially, we again found effects of stimulation on choice switching that were specific to task conditions: While tDCS did not differentially affect choice

switching in the specific condition ($t(53) = .61$, $p = .54$; Figure 4c), anodal tDCS increased and cathodal tDCS decreased choice switching in flexible trials ($t(53) = 3.47$, $p = .001$; Figure 4d; interaction between specific and flexible condition, $t(106) = 2.28$, $p = .025$). This pattern of results is therefore fully congruent with the model-based analysis reported above, which had shown increased or decreased switching between learning systems due to anodal or cathodal tDCS only in the flexible condition (Figure 3d). To more formally quantify this congruence, we computed correlations between the non-algorithmic choice switching and the indices of model preference and switching based on the learning algorithms. This revealed significant positive correlations between choice switching and both preference for MB learning ($R(108) = .46$, $p < .001$) and switching between learning systems ($R(108) = .21$, $p = .025$). Importantly, this relationship was only found for flexible trials (Figure 4b, right column) and was absent in specific trials (preference: $R(108) = -.02$, $p = .82$; switching: $R(108) = .08$, $p = .37$) (Figure 4b, left column). Since in flexible trials, our intervention only led to a change in the amount of switching between learning systems, the tDCS effect on choice switching may mainly reflect a stimulation-induced change in model arbitration behavior rather than a shift in the preference for one system over the other.

Figure 4. Alternative Non-algorithmic Analysis of Behavioral tDCS Effect



(a) *Choice switching (% of trials) between task conditions during the baseline session. Subjects switched between choice options from trial to trial more often in specific compared to flexible trials.*

(b) *Correlations between choice consistency in specific and flexible trials and parameters of model preference and switching. While there is no significant relationship in the specific condition (left column), choice switching is positively correlated with MB preference and the frequency of model switching within the flexible condition (right column).*

(c,d) *Differences in the tDCS-related change of choice switching between stimulation groups, defined as % of trials where the previous choice in S2 was not repeated. Direct current stimulation had a significant effect on choice switching, expressed as reduced switching from trial to trial after cathodal tDCS and relatively more choice switching after anodal stimulation only in the flexible condition (d).*

In all plots*, circles represent individual data points. If applicable, the central (colored) mark is the mean, the bottom and top of the shaded area represents the standard error of the mean (SEM).*

Discussion

Our results provide causal evidence that a prefrontal arbitration mechanism flexibly controls the use of MB or MF reinforcement learning. This mechanism is not only involved in changing which system is more dominant, but also appears to directly govern how much participants switch between these systems. Moreover, our results suggest that the stimulated mechanism controls both these aspects of arbitration based on the requirements of the current context: In the specific task condition, tDCS affected the preference for the MB learning system without any significant change in model switching, whereas in the flexible task condition, the stimulation significantly changed model switching without affecting the preference for MB over MF learning. The proposed neural arbitration mechanism therefore appears not to be deterministic but to flexibly adapt to environmental demands.

Context-dependent Adaptive Arbitration

The effects of tDCS varied across flexible and specific task conditions, suggesting that the arbitrator primarily controls the timing of switches between systems in flexible blocks, but

mainly affects the dominance between systems in specific blocks. This may reflect that the arbitrator indeed mainly exerts its effects via the system that is most suited for the current environment. As we could clearly show (see Figure 3a, b), the specific task condition favored MB learning whereas the flexible task conditions favored MF learning. In practice, however, MB learning can be used in both conditions, whereas MF learning is not very useful in specific trials. Hence, a reasonable strategy may be to maximally employ MB learning in the specific condition but also rely on the less effortful MF strategy in the flexible condition. Following this logic, the arbitrator may primarily favor the MB system in the specific task condition, but may selectively apply MB learning in the flexible condition only for trials where it is clearly superior, thereby increasing the switching between systems. This conjecture is consistent with the pattern of results we observe in both the learning model and the non-algorithmic analysis. Thus, our results support the notion that the targeted mechanism acts as a highly flexible, adaptive arbitrator between MB and MF learning that takes into account the predicted success associated with use of both learning systems in a given environment.

Possible Neural Mechanisms Underlying the Observed Effects

Since our present experiment only measured changes in behavior due to tDCS, one can only speculate about the exact neural mechanism underlying our observed effects on arbitration between MB and MF learning. At the computational level captured by the model we employed, anodal tDCS appeared to enhance the stability of arbitration control, by promoting a transition to the MB learning strategy whenever goal-directed control is needed or by encouraging flexibility to switch between MB and MF control. Cathodal tDCS appeared to have the opposite effect. One possible neural mechanism that may have been influenced by the tDCS and may have led to these changes is inhibitory control within a prefrontal cortex-striatal network. Lee and colleagues (2014) showed that the degree to

which behavior is guided by MB control is correlated with the strength of functional connectivity between the ventrolateral prefrontal cortex, the region targeted with tDCS in the current study, and the posterior putamen, the region found to encode the value information of the MF system (Tricomi et al., 2009). Moreover, MB behavior also correlated with connectivity between the posterior putamen and the ventromedial prefrontal cortex, a region found to encode chosen values which represent an integration of MB and MF value signals (Lee et al., 2014; Wunderlich et al., 2012a). Interestingly, the connectivity between those areas gets weaker during choices guided by MB control, suggesting an inhibitory mechanism (Lee et al., 2014). In summary, these results suggest that the putamen is not only crucial for guiding MF control but may also interact with the MB system via the ventrolateral and ventromedial prefrontal cortices. It may play a key role in a distributed neural network that is able to shift between MB and MF control, with the ventrolateral area possibly inhibiting the MF system via the striatum, thus ultimately acting as an arbitrator between both systems. It therefore appears plausible that the tDCS stimulation may have influenced switching between the two learning strategies by neuromodulation of the targeted ventrolateral structure, which may have changed its inhibitory control on MF value coding in the putamen and may thereby have shifted the balance towards MB control. This hypothesis emerging from our results could be directly tested in future studies combining tDCS with fMRI.

Clinical Relevance

Our results demonstrate that it is possible to shift participants towards MB as well as MF control with tDCS, depending on the polarity of the applied current. To our knowledge, this pattern of results is the first demonstration that both learning systems can be enhanced in their function by external interventions targeting neural processes. This is especially important as the extent of MB and MF learning have been linked to a wide range of

dysfunctional clinical conditions (Sebold et al., 2014; Voon et al., 2015) which are all associated with a lack in MB control. An intervention that enhances MB control, is easy to apply in a clinical setting (Brunoni et al., 2012; Fregni and Pascual-Leone, 2007), and only produces minor side effects (Poreisz et al., 2007), could be an important step towards the treatment of such patients.

A question that emerges in this context is why our results differ from those of previous experiments (Smittenaar et al., 2014) also targeting the prefrontal cortex with anodal tDCS. Closer inspection reveals multiple crucial differences between this study and our work, including the precise site of stimulation (left ventrolateral PFC area found by Lee and colleagues (2014) versus right dorsolateral PFC site associated with working memory performance by Feredoes and colleagues (2011)) as well as some procedural differences (current density, return electrode position, and individualized versus standardized EEG-position-based electrode localization). These differences will obviously have to be taken into account by future applied studies that may optimize the stimulation procedures for maximal behavioral effects.

Limitations of tDCS

While tDCS features many advantages in comparison to other stimulation methods, a notable disadvantage is its lack of spatial specificity (Bikson and Rahman, 2013). Additionally, the size and position of the electrodes can considerably alter the spatial pattern of electrical field in different areas (Wagner et al., 2007). Although we tried to mitigate those limitations with our montage, which features a limited area of possible current flow as well as a larger return electrode to limit stimulation effects underneath it, we cannot rule out that cortical areas adjacent to the targeted site may also have been affected. Any conclusions about the precise neural mechanisms bringing about the behavioral effects observed here thus await further investigation, for example with

combined tDCS-fMRI (see Hauser et al., 2016; Moisa et al., 2016).

Recent studies have also raised a debate about the general suitability of tDCS to generate measurable neurophysiological and cognitive effects across many different domains (Horvath et al., 2015a; Horvath et al., 2015b; Wiethoff et al., 2014; see also Polania et al., 2018). While these doubts are mitigated by many findings of reliable tDCS effects on decision making (Fecteau et al., 2007; Hecht et al., 2010; Knoch et al., 2007; Ruff et al., 2013), they underline that the parameters which stimulation is applied might play an important role in its outcome. Additionally, it has to be noted that tDCS might be best suited to alter ongoing neural activity (Wagner et al., 2007) as opposed to inducing neural activity in a bottom-up manner. tDCS outcomes thus do not only depend on technical details of the stimulation but also on the subject's cognitive and neural processing state induced by the paradigm. All these factors are crucial for understanding and comparing the effectiveness of direct current stimulation across different studies and domains.

Future Directions

Viewed from a broader perspective, our results may also contribute to the general understanding of neurostimulation effects on brain activity and its corresponding behavior. In our study, the exact same stimulation protocol caused very different observable behavioral effects depending on the cognitive state of subjects. More specifically, stimulation mainly affected the cognitive functions that had to be applied in certain task context (model preference versus model switching). For basic visuo-motor functions, it is already established that neural effects of stimulation on both the stimulated region and remote areas can vary strongly with both internal and external context factors (Bohning et al., 1999; Ruff et al., 2009, 2006). Our present results now illustrate that even high-level adaptive behavioral control processes may be affected differentially by the same stimulation protocol in different environmental contexts. This supports the notion that tDCS

is mainly modulating neural activity that is functionally relevant in a given context, rather than exerting static invariant effects on neural activity that affect behavior in a fixed way (Antal et al., 2014; also see, Woods et al., 2016). Future research combining tDCS with fMRI could further investigate the precise neural mechanisms underlying our behavioral results by examining how neural activity changes in the vlPFC and interconnected areas control the context-sensitive arbitration between MB and MF learning. In particular, further characterizing the possible inhibitory-control mechanism of the prefrontal-striatal network (as hypothesized above) could be an important direction for future studies.

Another question worth investigating is how the observed modification of MB and MF control may extend to additional contexts. Since the observed effects seem to differ between conditions within the same task, it is necessary to clarify the range by which arbitration between MB and MF learning can adapt to even more varied environmental demands or cognitive states. For example, different experiments used qualitatively different two-step tasks to measure MB and MF control (compare Daw et al., 2011 and Gläscher et al., 2010). Future work should therefore extend the context-sensitive nature of the neural mechanism identified here to these alternative paradigms, and possibly to novel tasks specifically designed to elicit the need for different types of balance between MB and MF control.

Funding

This work was supported by the Swiss National Science Foundation with the “Sinergia” grant “Neuroeconomics of value-based decision making” (141965) and the Samsung Research Funding Center of Samsung Electronics (SRFC-TC1603-06).

Acknowledgments

We want to thank Rafael Polania for assistance in applying the direct current stimulation.

References

Adams CD, Dickinson A. 1981. Instrumental responding following reinforcer devaluation. Q. J. Exp. Psychol. Sect. B. 33:109–121.

Antal A, Ambrus GG, Chaieb L. 2014. Toward unraveling reading-related modulations of tDCS-induced neuroplasticity in the human visual cortex. Front. Psychol. 5:642.

Baayen RH, Davidson DJ, Bates DM. 2008. Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. 59:390–412.

Balleine BW, Dickinson A. 1998. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology. 37:407–419.

Batsikadze G, Moliadze V, Paulus W, Kuo MF, Nitsche MA. 2013. Partially non-linear stimulation intensity-dependent effects of direct current stimulation on motor cortex excitability in humans. J. Physiol. 591:1987–2000.

Bikson M, Rahman A. 2013. Origins of specificity during tDCS: anatomical, activity-selective, and input-bias mechanisms. Front. Hum. Neurosci. 7:688.

Bohning DE, Shastri A, McConnell KA, Nahas Z, Lorberbaum JP, Roberts DR, Teneback C, Vincent DJ, George MS. 1999. A combined TMS/fMRI study of intensity-dependent TMS over motor cortex. Biol. Psychiatry 45:385–394.

Brainard DH. 1997. The psychophysics toolbox. Spat. Vis. 10:433–436.

Brunoni AR, Nitsche MA, Bolognini N, Bikson M, Wagner T, Merabet L, Edwards DJ, Valero-Cabre A, Rotenberg A, Pascual-Leone A, Ferrucci R, Priori A, Boggio PS, Fregni F. 2012. Clinical research with transcranial direct current stimulation (tDCS): challenges and future directions. Brain Stimul. 5:175–195.

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron. 69:1204–1215.

Daw ND, Niv Y, Dayan P. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. 8:1704–1711.

Dayan P, Niv Y. 2008. Reinforcement learning: the good, the bad and the ugly. Curr. Opin. Neurobiol. 18:185–196.

de Wit S, Barker RA, Dickinson AD, Cools R. 2011. Habitual versus goal-directed action control in Parkinson disease. J. Cogn. Neurosci. 23:1218–1229.

Deserno L, Huys QJM, Boehme R, Buchert R, Heinze HJ, Grace AA, Dolan RJ, Heinz A, Schlagenhauf F. 2015. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. Proc. Natl. Acad. Sci. U. S. A. 112:1595–1600.

Dolan RJ, Dayan P. 2013. Goals and Habits in the Brain. Neuron 80:312–325.

Everitt BJ, Robbins TW. 2005. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nat. Neurosci. 8:1481–1489.

Fecteau S, Pascual-Leone A, Zald DH, Liguori P, Théoret H, Boggio PS, Fregni F. 2007. Activation of prefrontal cortex by transcranial direct current stimulation reduces appetite for risk during ambiguous decision making. J. Neurosci. 27:6212–6218.

Feredoes E, Heinen K, Weiskopf N, Ruff C, Driver J. 2011. Causal evidence for frontal involvement in memory target maintenance by posterior brain areas during distracter interference of visual working memory. Proc. Natl. Acad. Sci. U. S. A. 108:17510–17515.

Fregni F, Pascual-Leone A. 2007. Technology insight: noninvasive brain stimulation in neurology-perspectives on the therapeutic potential of rTMS and tDCS. Nat. Clin. Pract. Neurol. 3:383–393.

Gevins A, Cutillo B. 1993. Spatiotemporal Dynamics of Component Processes in Human

Working-Memory. Electroencephalogr. Clin. Neurophysiol. 87:128–143.

Gillan CM, Otto AR, Phelps EA, Daw ND. 2015. Model-based learning protects against forming habits. Cogn. Affect. Behav. Neurosci. 15:523–536.

Gillan CM, Robbins TW. 2014. Goal-directed learning and obsessive-compulsive disorder. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 369:20130475.

Gläscher J, Daw N, Dayan P, O'Doherty JP. 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 66:585–595.

Gläscher JP, O'Doherty JP. 2010. Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. Wiley Interdiscip. Rev. Cogn. Sci. 1: 501–510.

Haruno M, Kawato M. 2006. Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. J. Neurophysiol. 95:948–959.

Hauser TU, Rütsche B, Wurmitzer K, Brem S, Ruff CC, Grabner RH. 2016. Neurocognitive effects of transcranial direct current stimulation in arithmetic learning and performance: a simultaneous tDCS-fMRI study. Brain Stimul. 9:850-858.

Hecht D, Walsh V, Lavidor M. 2010. Transcranial direct current stimulation facilitates decision making in a probabilistic guessing task. J. Neurosci. 30:4241–4245.

Horvath JC, Forte JD, Carter O. 2015a. Evidence that transcranial direct current stimulation (tDCS) generates little-to-no reliable neurophysiologic effect beyond MEP amplitude modulation in healthy human subjects: a systematic review. Neuropsychologia. 66:213-236.

Horvath JC, Forte JD, Carter O. 2015b. Quantitative review finds no evidence of cognitive effects in healthy populations from single-session transcranial direct current stimulation (tDCS). Brain Stimul. 8:535–550.

Knoch D, Nitsche MA, Fischbacher U, Eisenegger C, Pascual-Leone A, Fehr E. 2007. Studying the neurobiology of social interaction with transcranial direct current stimulation—the example of punishing unfairness. Cereb. Cortex. 18:1987-1990.

Lagarias JC, Reeds JA, Wright MH, Wright PE. 1998. Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. SIAM Journal on Optimization. 9:112–147.

Lee SW, Shimojo S, O'Doherty JP. 2014. Neural computations underlying arbitration between model-based and model-free learning. Neuron. 81:687–699.

Luce RD. 1959. Individual Choice Behavior. New York: Wiley.

Mars RB, Shea NJ, Kolling N, Rushworth MFS. 2012. Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. Q. J. Exp. Psychol. 65:252–267.

Moisa M, Polania R, Grueschow M, Ruff CC. 2016. Brain network mechanisms underlying motor enhancement by transcranial entrainment of gamma oscillations. J. Neurosci. 36:12053-12065.

Nitsche MA, Paulus W. 2001. Sustained excitability elevations induced by transcranial DC motor cortex stimulation in humans. Neurology. 57:1899–1901.

Nitsche MA, Cohen LG, Wassermann, EM, Priori A, Lang N, Antal A, Paulus W, Hummel F, Boggio PS, Fregni F, Pascual-Leone A. 2008. Transcranial direct current stimulation: State of the art 2008. Brain Stimul. 1:206–223.

Nitsche MA, Nitsche MS, Klein CC, Tergau F, Rothwell JC, Paulus W. 2003. Level of action of cathodal DC polarisation induced inhibition of the human motor cortex. Clin. Neurophysiol. 114:600–604.

Nitsche MA, Paulus W. 2000. Excitability changes induced in the human motor cortex by weak transcranial direct current stimulation. J. Physiol. 527:633–639.

O'Doherty JP, Hampton A, Kim H. 2007. Model-based fMRI and its application to reward

learning and decision making. Ann. N. Y. Acad. Sci. 1104:35–53.

O'Doherty JP, Lee SW, McNamee D. 2015. The structure of reinforcement-learning mechanisms in the human brain. Curr. Opin. Behav. Sci. 1:94–100.

Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND. 2013. Working-memory capacity protects model-based learning from stress. Proc. Natl. Acad. Sci. U. S. A. 110:20941–20946.

Polania R, Nitsche MA, Ruff CC. 2018. Studying and modifying brain function with non-invasive brain stimulation. Nat. Neurosci. 21:174–187.

Poreisz C, Boros K, Antal A, Paulus W. 2007. Safety aspects of transcranial direct current stimulation concerning healthy subjects and patients. Brain Res. Bull. 72:208–14.

Prévost C, McNamee D, Jessup RK, Bossaerts P, O'Doherty JP. 2013. Evidence for model-based computations in the human amygdala during Pavlovian conditioning. PLoS Comput. Biol. 9:e1002918.

Redgrave P, Rodriguez M, Smith Y, Rodriguez-Oroz MC, Lehericy S, Bergman H, Agid Y, DeLong MR, Obeso JA. 2010. Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. Nat. Rev. Neurosci. 11:760–772.

Ruff CC, Blankenburg F, Bjoertomt O, Bestmann S, Freeman E., Haynes JD, Rees G, Josephs O, Deichmann R, Driver J. 2006. Concurrent TMS-fMRI and psychophysics reveal frontal influences on human retinotopic visual cortex. Curr. Biol. 16:1479–1488.

Ruff CC, Driver J, Bestmann S. 2009. Combining TMS and fMRI: From "virtual lesions" to functional-network accounts of cognition. Cortex. 45:1043–1049.

Ruff CC, Ugazio G, Fehr E. 2013. Changing Social Norm Compliance with Noninvasive Brain Stimulation. Science. 342:482–484.

Sebold M, Deserno L, Nebe S, Schad DJ, Garbusow M, Hägele C, Keller J, Jünger E, Kathmann N, Smolka MN, Smolka M, Rapp, MA, Schlagenhauf F, Heinz A, Huys

QJM. 2014. Model-based and model-free decisions in alcohol dependence. Neuropsychobiology 70:122–131.

Smittenaar P, FitzGerald THB, Romei V, Wright ND, Dolan RJ. 2013. Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. Neuron. 80:914–919.

Smittenaar P, Prichard G, FitzGerald THB, Diedrichsen J, Dolan RJ. 2014. Transcranial direct current stimulation of right dorsolateral prefrontal cortex does not affect model-based or model-free reinforcement learning in humans. PLoS One. 9:e86850.

Sutton RS, Barto AG. 1998. Reinforcement Learning: An Introduction. Cambridge: MIT press.

Thorndike EL. 1933. A proof of the law of effect. Science. 77:173–175.

Tricomi E, Balleine BW, O'Doherty JP. 2009. A specific role for posterior dorsolateral striatum in human habit learning. Eur. J. Neurosci. 29:2225–2232.

Utz KS, Dimova V, Oppenländer K, Kerkhoff G. 2010. Electrified minds: Transcranial direct current stimulation (tDCS) and Galvanic Vestibular Stimulation (GVS) as methods of non-invasive brain stimulation in neuropsychology-A review of current data and future implications. Neuropsychologia 48:2789–2810.

Voon V, Derbyshire K, Rück C, Irvine MA, Worbe Y, Enander J, Schreiber LRN, Gillan C, Fineberg NA, Sahakian BJ, Robbins TW, Harrison NA, Wood J, Daw ND, Dayan P, Grant JE, Bullmore ET. 2015. Disorders of compulsivity: a common bias towards learning habits. Mol. Psychiatry. 20:345–352.

Wagner T, Fregni F, Fecteau S, Grodzinsky A, Zahn M, Pascual-Leone A. 2007. Transcranial direct current stimulation: a computer-based human model study. Neuroimage. 35:1113-1124.

Wiethoff S, Hamada M, Rothwell JC. 2014. Variability in response to transcranial direct

current stimulation of the motor cortex. Brain Stimul. 7:468-475.

Woods AJ, Antal A, Bikson M, Boggio PS, Brunoni AR, Celnik P, Cohen LG, Fregni F, Herrmann CS, Kappenman ES, Knotkova H, Liebetanz D, Miniussi C, Miranda PC, Paulus W, Priori A, Reato D, Stagg C, Wenderoth N, Nitsche MA. 2016. A technical guide to tDCS, and related non-invasive brain stimulation tools. Clin. Neurophysiol. 127:1031–1048.

Wunderlich K, Dayan P, Dolan RJ. 2012a. Mapping value based planning and extensively trained choice in the human brain. Nat. Neurosci. 15:786–791.

Wunderlich K, Smittenaar P, Dolan RJ. 2012b. Dopamine enhances model-based over model-free choice behavior. Neuron. 75:418–424.